УДК 519.25

А.П. Сарычев
Институт технической механики НАН Украины и НКА Украины, г. Днепропетровск
Sarychev@prognoz.dp.ua

Дискриминантный анализ по независимым признакам в условиях структурной неопределенности

Рассмотрена задача поиска дискриминантной функции оптимальной сложности в условиях независимых признаков. Рассмотрен способ сравнения дискриминантных функций, который основан на разбиении наблюдений на обучающие и проверочные подвыборки: обучающие подвыборки используются для оценивания коэффициентов дискриминантной функции, а проверочные подвыборки – для оценивания ее качества классификации. В данной статье общие результаты, полученные автором, применены для частного случая, когда признаки статистически независимы.

Решение задачи дискриминантного анализа в условиях структурной неопределенности – с определением оптимального множества признаков – предполагает принятие какого-либо способа сравнения дискриминантных функций, построенных на различных множествах признаков.

Популярным в приложениях является способ сравнения, основанный на разбиении наблюдений на обучающие и проверочные подвыборки: обучающие подвыборки используются для оценивания коэффициентов дискриминантной функции, а проверочные подвыборки – для оценивания ее качества классификации. Этот способ традиционно трактуется как эвристический прием, хотя факт существования для него оптимального множества признаков неоднократно подтвержден методом статистических испытаний [1-7]. В рамках метода группового учета аргументов (МГУА) для этого способа аналитически исследован критерий качества для сравнения дискриминантных функций, построенных на различных множествах признаков [8-11]. В данной статье результаты, полученные в [8-11], применены для частного случая, когда признаки статистически независимы.

Для решения задачи дискриминантного анализа в условиях структурной неопределенности кроме способа сравнения дискриминантных функций требуется указать алгоритм генерирования наборов признаков, включаемых в дискриминантную функцию. Предполагается, что в качестве такового принят полный перебор всех возможных сочетаний признаков.

Постановка задачи

Предположим, что

$$\mathrm{X}_k = [\mathrm{x}_{k1}, \mathrm{x}_{k2}, \ldots, \mathrm{x}_{kn_k}] \tag{1}$$

– выборка n_k независимых наблюдений m-мерного случайного вектора $з_k$ из генеральной совокупности P_k, имеющего m-мерное нормальное распределение с

неизвестным математическим ожиданием ч_k и неизвестной невырожденной ковариационной матрицей У_X

$$\text{з}_k \sim N_m(\text{ч}_k, \text{У}_X), \qquad (2)$$

где k – номер I или II генеральных совокупностей P_I и P_{II}, причем $\text{ч}_I \neq \text{ч}_{II}$.

Согласно предположению (2) для наблюдений (1) выполняется

$$X_{ki} = \text{ч}_k + \text{о}_{ki}, \quad k=I,II, \quad i=1,\ 2,...,n_k, \qquad (3)$$

где $\text{о}_{ki} \sim N_m(0_m, \text{У}_X)$ – независимые случайные векторы, распределенные по m-мерному нормальному закону с нулевым математическим ожиданием и ковариационной матрицей У_X

$$E\{\text{о}_{ki}\}=0_m; \quad E\{\text{о}_{ki}\text{о}_{ki}^T\}=\text{У}_X; \quad i=1,\ 2,...,n_k; \quad k=I,II; \qquad (4)$$

$$E\{\text{о}_{ki_1}\text{о}_{ki_2}^T\}=O_{(m\text{ч}m)}; \quad i_1,i_2=1,\ 2,...,n_k; \quad i_1\text{№}_2; \qquad (5)$$

$$E\{\text{о}_{Ii_1}\text{о}_{IIi_2}^T\}=O_{(m\text{ч}m)}; \quad i_1,i_2=1,\ 2,...,n_k, \qquad (6)$$

где 0_m – нулевой m-мерный вектор-столбец и $O_{(m\text{ч}m)}$ – нулевая $(m \times m)$-матрица, $E\{\cdot\}$ – знак математического ожидания.

Предположим, что априорные вероятности появления наблюдений из P_I и P_{II} известны и соответственно равны p_I и p_{II}, причем $p_I + p_{II} = 1$.

Предположим, что введены цены ошибочных классификаций: $c(I/II)$ – цена ошибочной классификации наблюдения из совокупности P_{II} в качестве наблюдения из P_I, а $c(II/I)$ – цена ошибочной классификации наблюдения из совокупности P_I в качестве наблюдения из P_{II}. Правильная классификация не оценивается.

Линейная дискриминантная функция, при которой в указанных предположениях ожидаемая ошибка классификации минимальна, имеет вид

$$R(x)=x^T\text{д}-0{,}5(\text{ч}_I+\text{ч}_{II})^T\text{д}-\ln c_0, \qquad (7)$$

где параметр c_0 определяется ценами ошибочных классификаций и априорными вероятностями появления наблюдений

$$c_0 = c(I/II)\pi_{II}(c(II/I)\pi_I)^{-1}, \qquad (8)$$

а коэффициенты дискриминантной функции д определяются параметрами генеральных совокупностей

$$\text{д}=\text{У}_X^{-1}(\text{ч}_I-\text{ч}_{II}). \qquad (9)$$

Решающее правило формулируется следующим образом: если для наблюдения x^* выполняется $R(x^*) \geq 0$, то $x^* \in P_I$ – оно относится к первой группе; если $R(x^*) < 0$, то $x^* \in P_{II}$ – оно относится ко второй группе.

Получение по выборочным наблюдениям (1) – (3) оценок коэффициентов д в дискриминантной функции (7) с последующим статистическим анализом является задачей дискриминантного анализа, поставленной **в узком смысле**.

Фишеровской оценкой д является оценка

$$\hat{d}=S^{-1}(\tilde{x}_I - \tilde{x}_{II}), \qquad (10)$$

где $\tilde{x}_{I}$ и $\tilde{x}_{II}$ – оценки математических ожиданий $ч_I$ и $ч_{II}$

$$\tilde{x}_{I}=(n_I)^{-1}\sum_{i=1}^{n_I} X_{Ii}, \qquad \tilde{x}_{II}=(n_{II})^{-1}\sum_{i=1}^{n_{II}} X_{IIi}\,; \tag{11}$$

матрица S – оценка ковариационной матрицы $У_X$

$$S=(n_I+n_{II}-2)^{-1}[x_I x_I^T+x_{II}x_{II}^T], \tag{12}$$

а x_I и x_{II} – $(m\times n_I)$- и $(m\times n_{II})$-матрицы, составленные из отклонений наблюдений (1) и (3) от оценок $\tilde{x}_{I}$ и $\tilde{x}_{II}$ соответственно

$$x_I=[X_{I,1}-\tilde{x}_{I},X_{I,2}-\tilde{x}_{I},...,X_{I,n_I}-\tilde{x}_{I}], \tag{13}$$

$$x_{II}=[X_{II,1}-\tilde{x}_{II},X_{II,2}-\tilde{x}_{II},...,X_{II,n_{II}}-\tilde{x}_{II}]. \tag{14}$$

Оценка (10) является решением задачи максимизации функционала

$$Ц=\frac{d^T(\tilde{x}_{I}-\tilde{x}_{II})(\tilde{x}_{I}-\tilde{x}_{II})^T d}{d^T S\ d} ® \max_d \tag{15}$$

при ограничении

$$\frac{(\tilde{x}_{I}-\tilde{x}_{II})^T d}{d^T S\ d}=1. \tag{16}$$

Значение функционала (15) при оптимальном значении $\hat{d}=S^{-1}(\tilde{x}_{I}-\tilde{x}_{II})$ –

$$D^2=(\tilde{x}_{I}-\tilde{x}_{II})^T S^{-1}(\tilde{x}_{I}-\tilde{x}_{II}) \tag{17}$$

является выборочной оценкой расстояния Махаланобиса

$$ф_X^2=(ч_I-ч_{II})^T У_X^{-1}(ч_I-ч_{II}). \tag{18}$$

Для математического ожидания величины D^2 выполняется:

$$E\{D^2\}=\frac{r}{r-m-1}\tau_X^2+m\,c^{-1}, \tag{19}$$

где $r=n_I+n_{II}-2,\ \ c^{-1}=(n_I^{-1}+n_{II}^{-1})$.

Пусть X – множество m компонент векторов $з_I$ и $з_{II}$, над которыми проведены наблюдения; $\overset{o}{X}$ – множество $\overset{o}{m}$ ($1\le\overset{o}{m}<m$) компонент векторов $з_I$ и $з_{II}$, для математических ожиданий которых выполнено: $\overset{o}{ч}_{Ij}\neq\overset{o}{ч}_{IIj},\ \ j=1,2,...,\overset{o}{m},$ причем множество $\overset{o}{X}$ неизвестно, известно лишь, что $\overset{o}{X}\subseteq X$.

Если априорно неизвестно, какие именно компоненты из множества X следует включать в дискриминантную функцию, то говорят о задаче дискриминантного анализа, поставленной в широком смысле. В этом случае по наблюдениям X_I и X_{II} требуется определить множество компонент, которые необходимо включать в дискриминантную функцию, и оценить коэффициенты линейной дискриминантной функции в пространстве этих компонент.

Известно [5-7], что оценка расстояния Махаланобиса (17) не может быть использована для решения задачи дискриминантного анализа, поставленной в широком смысле, поскольку эта оценка увеличивается при добавлении новых компонент в дискриминантную функцию (в решающее правило). Распространенным приемом в этом случае является назначение «порога» для величины приращения оценки расстояния Махаланобиса (17) при добавлении новой компоненты в дискриминантную функцию [12], [13]. Но такое назначение привносит субъективность в получаемое решение.

Воспользуемся результатами работ [8-11] при дополнительном предположении о том, что признаки статистически независимы, т.е. ковариационная матрица $У_X$ в (2), (4), (9), (18) имеет диагональный вид.

1. Критерий качества дискриминантных функций для способа с разбиением наблюдений на обучающие и проверочные подвыборки

Пусть на этапе с номером $s\ (s = 1,2,..., m)$ алгоритма полного перебора сочетаний признаков в дискриминантную функцию может быть включено только s компонент из множества X_k, составляющих текущее анализируемое множество V. Пусть множеству компонент V соответствуют: 1) V_I и V_{II} – $(s \times n_I)$- и $(s \times n_{II})$-матрицы наблюдений из генеральных совокупностей P_I и P_{II}; 2) $н_I$ и $н_{II}$ – $(s \times 1)$-векторы математических ожиданий для наблюдений из P_I и P_{II}; 3) $У_V$ – ковариационная $(s \times s)$-матрица. Рассмотрим оценку расстояния Махаланобиса, рассчитываемую с учетом разбиения наблюдений на обучающие и проверочные подвыборки

$$D^2_{AB}(V) = \frac{\hat{d}_A^T\,(\tilde{v}_{IB} - \tilde{v}_{IIB})\,(\tilde{v}_{IB} - \tilde{v}_{IIB})^T\,\hat{d}_A}{\hat{d}_A^T\, S_B\, \hat{d}_A}. \quad (20)$$

В (20) (sЧ1)-вектор $\hat{d}_A$ представляет собой рассчитанную на подвыборке A фишеровскую оценку коэффициентов дискриминантной функции, которая построена в пространстве компонент множества V

$$\hat{d}_A = S_A^{-1}\,(\tilde{v}_{IA} - \tilde{v}_{IIA}), \quad (21)$$

где $(s \times 1)$-векторы $\tilde{v}_{IA}$ и $\tilde{v}_{IIA}$ – оценки математических ожиданий $\mathbf{н}_I$ и $\mathbf{н}_{II}$

$$\tilde{v}_{kA} = (n_{kA})^{-1}\sum_{i=1}^{n_{kA}} V_{kiA}, \quad k = I,II; \quad (22)$$

$(s \times s)$-матрица S_A – несмещенная оценка ковариационной матрицы $У_V$

$$S_A = (n_{IA} - n_{IIA} - 2)^{-1}\ [v_{IA}v_{IA}^T + v_{IIA}v_{IIA}^T]. \quad (23)$$

В (23) v_{kA} (k=I,II) – $(s \times n_k)$-матрицы, составленные из отклонений наблюдений V_{kA} компонент множества V от оценок $\tilde{v}_{kA}$

$$v_{kA} = [V_{k1A} - \tilde{v}_{kA}, V_{k2A} - \tilde{v}_{kA}, ..., V_{kn_kA} - \tilde{v}_{kA}]. \quad (24)$$

В (20) $(s \times 1)$-векторы $\tilde{v}_{IB}$ и $\tilde{v}_{IIB}$ вычисляются аналогично (22), а $(s \times s)$-матрица S_B – аналогично (23) – (24); n_{IA} и n_{IIA}, n_{IB} и n_{IIB} – объемы обучающих и проверочных подвыборок соответственно такие, что выполняется

$$n_{IA} + n_{IB} = n_I \text{ и } n_{IIA} + n_{IIB} = n_{II}.$$

Используя (21), для (20) получаем

$$D^2_{AB}(V) = \frac{(\tilde{v}_{IA} - \tilde{v}_{IIA})^T S_A^{-1} (\tilde{v}_{IB} - \tilde{v}_{IIB})(\tilde{v}_{IB} - \tilde{v}_{IIB})^T S_A^{-1} (\tilde{v}_{IA} - \tilde{v}_{IIA})}{(\tilde{v}_{IA} - \tilde{v}_{IIA})^T S_A^{-1}\, S_B\, S_A^{-1} (\tilde{v}_{IA} - \tilde{v}_{IIA})}. \quad (25)$$

Для математического ожидания случайной величины (25) с учетом результатов [9], [11] выполняется

$$E\{D^2_{AB}(V)\} = \left(\tau_V^2 - \frac{\tau_V^2\ [s-(r_1-1)\ /\ (r_1-s)]\ c_A^{-1}}{\tau_V^2 + s\ c_A^{-1}} + c_B^{-1}\ \frac{r_1 - 1}{r_1 - s} \right) \frac{r_1 - s}{r_1 - 1}, \quad (26)$$

где

$$r_1 = n_{IA} + n_{IIA} - 2, \quad c_A^{-1} = (n_{IA}^{-1} + n_{IIA}^{-1}), \quad c_B^{-1} = (n_{IB}^{-1} + n_{IIB}^{-1}); \quad (27)$$

$ф_V^2 = (н_I - н_{II})^T У_V^{-1} (н_I - н_{II})$ – расстояние Махаланобиса для множества компонент V.

2. Исследование математического ожидания критерия качества дискриминантной функции в зависимости от состава признаков

Покажем, что критерий качества дискриминантной функции (20) позволяет решать задачу дискриминантного анализа, поставленную в широком смысле. Покажем существование оптимального множества признаков для способа с разбиением наблюдений на обучающие и проверочные подвыборки и сформулируем условия, при которых оптимальная дискриминантная функция упрощается по числу входящих в нее признаков. С этой целью исследуем зависимость величины $D^2_{AB}(V)$ от состава множества V.

Определение 1. *Оптимальным множеством признаков называется множество V_{OPT}:*

$$V_{OPT} = \arg\max_{V \subseteq X} E\{D^2_{AB}(V)\}. \quad (28)$$

Определение 2. *Оптимальной по количеству и составу признаков называется фишеровская дискриминантная функция, построенная на множестве V_{OPT}.*

Разобьем множество статистически независимых компонент X на непересекающиеся подмножества

$$X = \overset{o}{X} \cup \tilde{R}, \quad (29)$$

где $\overset{o}{X} \neq \varnothing$ ($\varnothing$ – пустое множество) – множество компонент ($\overset{o}{m}$ – их число), для математических ожиданий которых выполнено $\overset{o}{ч}_{Ih} \neq \overset{o}{ч}_{IIh}, h = 1, 2, ..., \overset{o}{m}$; $\tilde{R}$ – множество компонент, для математических ожиданий которых выполнено $\tilde{c}_{Ih} = \tilde{c}_{IIh}$, $h = 1, 2, ..., \tilde{l}$ где $\tilde{l}$ – их число.

Рассмотрим величину

$$Д(V)=E\{D^2_{AB}(\overset{\circ}{X})\}-E\{D^2_{AB}(V)\} \tag{30}$$

для двух случаев состава множества компонент V .

Пусть в первом случае множество V таково, что выполняется: $V=\overset{\circ}{X}\bigcup\tilde{r}$, где $\tilde{r}\in\tilde{R}$ (в дискриминантную функцию включен избыточный признак).

Учитывая (26) – (27), для величины $Д_1(V)$ получаем

$$\begin{aligned}Д_1(V)&=E\{D^2_{AB}(\overset{\circ}{X})\}-E\{D^2_{AB}(\overset{\circ}{X}\bigcup\tilde{r})\}=\\&=\left(\tau^2_{\overset{\circ}{X}}-\frac{\tau^2_{\overset{\circ}{X}}\,[\overset{\circ}{m}-(r_1-1)/(r_1-\overset{\circ}{m})]\;c_A^{-1}}{\tau^2_{\overset{\circ}{X}}+\overset{\circ}{m}\,c_A^{-1}}+c_B^{-1}\;\frac{r_1-1}{r_1-\overset{\circ}{m}}\right)\frac{r_1-\overset{\circ}{m}}{r_1-1}-\\&-\left(\tau^2_V-\frac{\tau^2_V\,[(\overset{\circ}{m}+1)-(r_1-1)/(r_1-\overset{\circ}{m}-1)]\;c_A^{-1}}{\tau^2_V+\overset{\circ}{m}\,c_A^{-1}}+c_B^{-1}\;\frac{r_1-1}{r_1-\overset{\circ}{m}-1}\right)\frac{r_1-\overset{\circ}{m}-1}{r_1-1}.\end{aligned} \tag{31}$$

В соответствии с леммой 2 в работах [8-10] для расстояний Махаланобиса введенных в (29) множеств V и $\overset{\circ}{X}$ выполняется: $\tau^2_V=\tau^2_{\overset{\circ}{X}}$.

С учетом этого, ограничившись точностью $(1/r_1)$, пренебрегая членами порядка $(1/r_1^2)$, получаем

$$Д_1(V)=\frac{1}{r_1-1}\cdot\left\{\frac{(\tau^2_{\overset{\circ}{X}})^3+(\tau^2_{\overset{\circ}{X}})^2\,r_1c_A^{-1}}{(\tau^2_{\overset{\circ}{X}}+\overset{\circ}{m}\,c_A^{-1})(\tau^2_{\overset{\circ}{X}}+(\overset{\circ}{m}+1)c_A^{-1})}\right\}>0. \tag{32}$$

Из положительной определенности величины $Д_1(V)$ в (32) следует, что добавление признака $\tilde{r}\in\tilde{R}$ не улучшает качество дискриминантной функции по рассматриваемому критерию.

Пусть во втором случае множество V таково, что выполняется: $\overset{\circ}{X}=V\bigcup\overset{\circ}{x}$, где $\overset{\circ}{x}\in\overset{\circ}{X}$ (в дискриминантной функции пропущен один признак).

Учитывая (26) – (27), для величины $Д_2(V)$ получаем

$$\begin{aligned}Д_2(V)&=E\{D^2_{AB}(\overset{\circ}{X})\}-E\{D^2_{AB}(V)\}=\\&=\left(\tau^2_{\overset{\circ}{X}}-\frac{\tau^2_{\overset{\circ}{X}}\,[\overset{\circ}{m}-(r_1-1)/(r_1-\overset{\circ}{m})]\;c_A^{-1}}{\tau^2_{\overset{\circ}{X}}+\overset{\circ}{m}\,c_A^{-1}}+c_B^{-1}\;\frac{r_1-1}{r_1-\overset{\circ}{m}}\right)\frac{r_1-\overset{\circ}{m}}{r_1-1}-\\&-\left(\tau^2_V-\frac{\tau^2_V\,[(\overset{\circ}{m}-1)-(r_1-1)/(r_1-\overset{\circ}{m}+1)]\;c_A^{-1}}{\tau^2_V+(\overset{\circ}{m}-1)\,c_A^{-1}}+c_B^{-1}\;\frac{r_1-1}{r_1-\overset{\circ}{m}+1}\right)\frac{r_1-\overset{\circ}{m}+1}{r_1-1}.\end{aligned} \tag{33}$$

В соответствии с леммой 3 в работах [8-10] для расстояний Махаланобиса введенных в (29) множеств V и $\overset{o}{X}$ выполняется соотношение: $\tau_V^2 = \tau^2_{\overset{o}{X}} - \gamma^2$, где $\gamma^2 > 0$ – составляющая расстояния Махаланобиса, обусловленная пропущенным признаком.

С учетом этого, ограничившись точностью $(1/r_1)$, пренебрегая членами порядка $(1/r_1^2)$, получаем

$$Д_2(V) = \frac{1}{(r_1-1)(\tau^2_{\overset{o}{X}} + \overset{o}{m} c_A^{-1})[(\tau^2_{\overset{o}{X}} - \gamma^2) + (\overset{o}{m}-1)c_A^{-1}]} \cdot \Big\{ -\tau^2_{\overset{o}{X}} \tau^2_{\overset{o}{X}} (\tau^2_{\overset{o}{X}} - \gamma^2) +$$

$$+ \gamma^2 \tau^2_{\overset{o}{X}} (r_1 - \overset{o}{m} - 1)(\tau^2_{\overset{o}{X}} - \gamma^2) + \gamma^2 (r_1 - \overset{o}{m}) \overset{o}{m} c_A^{-1} (\tau^2_{\overset{o}{X}} - \gamma^2) + \gamma^2 (r_1 - \overset{o}{m} - 1) \tau^2_{\overset{o}{X}} (\overset{o}{m} - 1) c_A^{-1} \Big\}. \tag{34}$$

Величина $Д_2(V)$ может быть как положительной, так и отрицательной.

Если величина $Д_2(V)$ положительна, то признак $\overset{o}{x}$ необходимо включать в дискриминантную функцию. Если величина $Д_2(V)$ отрицательна, то признак $\overset{o}{x}$ не следует включать в дискриминантную функцию, поскольку это включение приведет к уменьшению критерия D_{AB}^2, т.е. добавление признака $\overset{o}{x} \in \overset{o}{X}$ не улучшает качество дискриминантной функции по рассматриваемому критерию.

Получим условие отрицательности $Д_2(V)$. Величина $Д_2(V)$ отрицательна, если отрицателен числитель (34). Поделив числитель на положительную величину $\tau^2_{\overset{o}{X}} (\tau^2_{\overset{o}{X}} - \gamma^2)$, для условия отрицательности $Д_2(V)$ получаем:

$$\gamma^2 \left[(r_1 - \overset{o}{m} - 1) + \frac{(r_1 - \overset{o}{m}) \overset{o}{m} c_A^{-1}}{\tau^2_{\overset{o}{X}}} + \frac{(r_1 - \overset{o}{m} - 1)(\overset{o}{m} - 1) c_A^{-1}}{\tau^2_{\overset{o}{X}} - \gamma^2} \right] < \tau^2_{\overset{o}{X}}. \tag{35}$$

Условие (35) представляет собой условие редукции (упрощения) дискриминантной функции, оптимальной по количеству и составу признаков.

Это условие можно упростить, если принять: $\tau^2_{\overset{o}{X}} = 12$, $\overset{o}{m} = 6$, а для $r_1 = n_{IA} + n_{IIA} - 2$ и $c_A^{-1} = (n_{IA}^{-1} + n_{IIA}^{-1})$ принять, что $n_{IA} \approx n_{IIA}$, $n_{IA} + n_{IIA} = n_A$. При выполнении этих предположений второе и третье слагаемые в квадратных скобках (35) приближенно равны единице, и условие (35) принимает вид:

$$\gamma^2 < \frac{\tau^2_{\overset{o}{X}}}{(n_A - \overset{o}{m} - 1)}. \tag{36}$$

Опыт практических применений и тестовые исследования способа с разбиением наблюдений на обучающие и проверочные подвыборки, проведенные методом статистических испытаний [1-7], показывают: 1) с увеличением объема выборок

наблюдений увеличивается количество признаков во множестве, на котором достигается наилучшее качество распознавания, а с уменьшением объема выборок наблюдений количество признаков в таком множестве уменьшается; 2) с увеличением расстояния Махаланобиса τ_X^2 между генеральными совокупностями (из которых получены выборки наблюдений) увеличивается количество признаков во множестве, на котором достигается наилучшее качество распознавания, а с уменьшением этого расстояния количество признаков в таком множестве уменьшается.

Полученные в статье результаты объясняют эти эмпирически установленные закономерности.

Отметим, что в асимптотике при $n_A \to \infty$ $(r_1 \to \infty)$ условие редукции не выполняется, поскольку $\text{г}^2 > 0$, т.е. $V_{OPT} = \overset{o}{X}$.

Выводы

В статье исследован и аналитически обоснован способ сравнения дискриминантных функций, основанный на разбиении выборок наблюдений на обучающие и проверочные подвыборки в случае статистически независимых признаков. Несмотря на успешное применение этого способа на практике и неоднократное подтверждение его работоспособности методом статистических испытаний, он традиционно считался эвристическим приёмом.

Для этого способа получено условие существования оптимального множества признаков, зависящих от параметров генеральных совокупностей и объемов выборок, и выявлены закономерности упрощения оптимальной дискриминантной функции при уменьшении объемов выборок и при уменьшении расстояния Махаланобиса между генеральными совокупностями.

Показано, что в условиях структурной неопределенности и отсутствия априорных оценок дисперсий независимых признаков исследованный способ позволяет решать задачу поиска дискриминантной функции оптимальной сложности.

Литература

1. Лбов Г.С. Выбор эффективной системы зависимых признаков // Вычислительные системы. – 1965. – Вып. 19. – С. 21–34.
2. Раудис Ш.Ю. Влияние объема выборки на точность выбора модели в задаче распознавания образов // Статистические проблемы управления. – 1981. – Вып. 50. – С. 9-30.
3. Раудис Ш. Влияние объема выборки на качество классификации (обзор) // Статистические проблемы управления. – Вильнюс: Ин-т мат. и киб. АН Литвы, 1984. – Вып. 66. – С. 9-42.
4. Распознавание образов: Состояние и перспективы / К. Верхаген, Р. Дейн, Ф. Грун и др. – Пер. с англ. – М.: Радио и связь, 1985. – 104 с.
5. Фомин Я.А., Тарловский Г.Р. Статистическая теория распознавания образов. – М.: Радио и связь, 1986. – 264 с.
6. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др.– М.: Финансы и статистика, 1989. – 608 с.
7. Проблемы построения систем анализа данных // Статистические проблемы управления / Ред. Ш. Раудис. – Вильнюс: Институт математики и кибернетики АН Литвы, 1990. – Вып. 93. – 247 с.
8. Мирошниченко Л.В., Сарычев А.П. Схема скользящего экзамена для поиска оптимального множества признаков в задаче дискриминантного анализа // Автоматика. – 1992. – № 1. – С. 35-44.
9. Сарычев А.П. Схема поиска оптимального множества признаков в задаче дискриминантного анализа с разбиением наблюдений на обучающие и проверочные подвыборки // Автоматика. – 1992. – № 4. – С. 39-43.

10. Сарычев А.П. Схема скользящего экзамена для определения оптимального множества признаков в задаче дискриминантного анализа // Проблемы управления и информатики. – 2002. – № 6. – С. 65-77.
11. Сарычев А.П. Решение задачи дискриминантного анализа в условиях структурной неопределенности на основе метода группового учета аргументов // Проблемы управления и информатики. – 2008. – № 3. – С. 100-112.
12. Дженнрич Р.И. Пошаговый дискриминантный анализ // Статистические методы для ЭВМ: Пер. с англ. – М.: Наука, 1986. – С. 94-112.
13. Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа: Пер. с англ. – М.: Наука, 1986. – 464 с.

О.П. Саричев

Дискримінантний аналіз за незалежними ознаками в умовах структурної невизначеності

Розглянуто задачу пошуку дискримінантної функції оптимальної складності в умовах незалежних ознак. Розглянуто спосіб порівняння дискримінантних функцій, що заснований на розбивці спостережень на навчальні і перевірні підвибірки: навчальні підвибірки використовуються для оцінювання коефіцієнтів дискримінантної функції, а перевірні підвибірки – для оцінювання її якості класифікації. У даній статті загальні результати, які отримані автором, застосовані для часткового випадку, коли ознаки статистично незалежні.

A.P. Sarychev

The Discriminant Analysis with Independent Futures in Structural Uncertainty Conditions

The problem of search of a discriminant function of optimum complexity in conditions of independent futures is considered. Comparison of quality of discriminant functions is based on partition of initial sample of the data on learning and checking parts: on learning subsample the coefficients of discriminant functions estimate, and on checking subsample the quality of discriminant functions estimate. In the given article the general rezults obtained by the author, are applied for a particular case, when the futures are statistically independent.

Статья поступила в редакцию 21.07.2008.